

Lecture 9

Fairness and Ethics

Dr Cillian McHugh

PS4168: Economic Psychology

Overview

- Reciprocity
- Public Goods
- Third Party Punishment
- Moral Judgment

Recap on last week!

- 3 'components' of mental accounting?
- Principles of hedonic framing?
- Examples of mental accounting in your lives?

Recap on last week!

- 3 'components' of mental accounting:
 - Evaluating
 - Categorisation
 - Balancing
- Principles of hedonic framing
 - segregate gains
 - integrate losses
 - integrate smaller losses with larger gains
 - segregate small gains from larger losses

Fairness and Ethics

Reciprocity

- **Reciprocity:** *like with like* responding
 - In response to friendly actions, people are frequently nicer and more cooperative
 - In response to hostile actions they are frequently much more nasty and even brutal
 - Much more so than predicted by the self-interest model
- '*A man ought to be a friend to his friend and repay gift with gift. People should meet smiles with smiles and lies with treachery*' (Fehr & Gächter, 2003, p. 510)
- Cooperative reciprocal tendencies: ***positive reciprocity***
- Retaliatory tendencies: ***negative reciprocity***

Reciprocity (contd.)

- Reciprocity \neq cooperative/retaliatory behaviour in repeated interactions
 - These behaviors arise because actors expect future material benefits from their actions
- For reciprocity, the actor is responding to friendly or hostile actions ***even if no material gains can be expected***
- Reciprocity is also fundamentally different from altruism
 - Altruism is a form of unconditional kindness
 - altruism given does not emerge as a response to altruism received
- Reciprocity is an in-kind response to beneficial or harmful acts (Fehr & Gächter, 2003, p. 511)

Reciprocity: Examples and Evidence

- *In groups* Identify as many examples of reciprocity as you can think of

Studying Reciprocity

Ultimatum/Dictator Games

- Player 1 (Proposer) may divide a sum of money (e.g., €100) with Player 2 (Responder)
- Player 2 can accept or reject the offer
 - If Player 2 rejects, **both** players get **nothing**

Ultimatum/Dictator Games: Results

- If the Proposer offers < 30% the Responder is highly likely to reject (Camerer & Thaler, 1995; Fehr & Gächter, 2003; Güth, Schmittberger, & Schwarze, 1982; Roth, 1995)
- Observed in various countries: Europe, USA, Asia (Fehr & Gächter, 2003; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991)
 - (Although seemingly not in the Peruvian Amazon, Henrich, 2000)
- Minimal influence of higher stakes
 - Three months income
 - Student sample: \$100 or more (L. A. Cameron, 1999; Fehr & Gächter, 2003; Hoffman, McCabe, & Smith, 1995; Slonim & Roth, 1998)

Trust/Gift Exchange Games

- A Proposer receives an amount of money x from the experimenter
 - Sends amount y to the Responder
 - $0 \leq y \leq x$
- The experimenter then triples the amount sent
 - Responder has $3y$
 - The Responder is then free to return a value z between 0 and $3y$ to the Proposer
 - $0 \leq z \leq 3y$

Trust/Gift Exchange Results

- Many Proposers send money and many Responders give back some money
- Frequently a positive correlation between y and z the amount sent back
 - At both the individual as well as at the aggregate level(Berg, Dickhaut, & McCabe, 1995; Fehr, Kirchsteiger, & Riedl, 1993; McCabe, Rassenti, & Smith, 1996)
- As with negative reciprocity, positive reciprocity does not appear to be affected by higher/lower stakes
 - Fehr & Tougareva (1995) provided subjects with ≈ 10 weeks wages
 - for a 2 hour study

Fairness vs Selfishness

- 40% – 66% of people behave reciprocally in one-shot situations
 - Showing concern for fairness (reciprocal types)
- ***but***
- 20% – 30 of participants do not reciprocate
 - Behave completely selfishly (self interested types)
 - *Selfish Minority*(Abbink, Irlenbusch, & Renner, 2000; Berg et al., 1995; Fehr & Falk, 1999; Gächter & Falk, 1999)

Fairness vs Selfishness: Causes

Fairness vs Selfishness: Causes

- Linked with testosterone in males (Burnham, 2007)
- Desire to maintain equity: *Inequality Aversion* (Bolton & Ockenfels, 2006)
- Desire to punish hostile intentions and to reward kind intentions (Blount, 1995; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 1999; Rabin, 1993)
- Type of person (Levine, 1998)
- Boundedly rational (Gale, Binmore, & Samuelson, 1995; Roth & Erev, 1995)



Fairness vs Selfishness: Causes

- No agreement on cause
 - But agreement that reciprocity is **stable**
- Implications and applications of reciprocity:
 - Discuss in groups.

Implications and Applications

- How does reciprocity change the nature of collective action?
(e.g., in firms, public bureaucracies, markets and the political sphere?)
- Can reciprocal people constrain the opportunistic tendencies of selfish people?
- When (and where) does reciprocity thrive over selfishness?
(and vice-versa?)
- Do explicit incentives crowd out or enhance voluntary cooperation?

Public Goods

Public Goods

- How should public goods be provided?
- Self interest dictates that everyone wants to 'free-ride' on the efforts of others

Public Goods

- How should public goods be provided?
- Self interest dictates that everyone wants to 'free-ride' on the efforts of others



Public Goods in the Lab

- In groups: Use your knowledge of the topics we've covered this semester to design a "Public Goods" Game

Public Goods in the Lab

- Fehr & Gächter (2000)
 - 4 group members
 - 20 tokens each
- Tokens may be *kept* or *invested*
 - *Keeping* 1 token earns 1 token
 - An *invested* token earns 0.4 of a token for each player
 - Private return on *investing* = 0.4 tokens
 - Social return on *investing* = 1.6 tokens

Public Goods in the Lab (contd.)

- If everyone *Keeps everything*:
 - Everyone earns 20 tokens
- If everyone *invests everything*:
 - Everyone earns 32 tokens ($4 \times 0.4 \times 20$)
- But it is in everyone's (private) interest to invest nothing and 'free ride'
 - *Keeping* is ***strictly dominant***



Public Goods and Reciprocity

- Positive reciprocity implies that people are willing to contribute something to the public good *if* others are also willing to contribute
- Contribution to the public good represents a kind action
 - Which induces reciprocally motivated people to contribute too (Fehr & Gächter, 2003; Sugden, 1984)
- **But**
 - *Selfish Minority*(Abbink et al., 2000; Berg et al., 1995; Fehr & Falk, 1999; Gächter & Falk, 1999)
 - 20% – 30 of participants do not reciprocate
 - Can a positive level of contributions to the public good can be sustained as an equilibrium?

Public Goods and Reciprocity

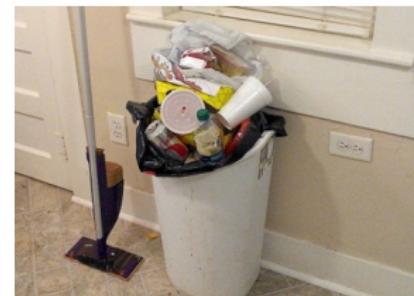
- The game used by Fehr & Gächter (2000) did not have any opportunity for direct retaliation in response to observed free riding
- Indirect negative reciprocity is possible
 - If subjects people interpret others' free riding that as a hostile act
 - They can "punish" others by also free riding
- Different motivations, but everyone ends up free riding
 - Self-interested types choose to free ride because they are self-interested
 - Reciprocal types free-ride because they observe others free-riding
- Examples?

Reciprocal Free Riding

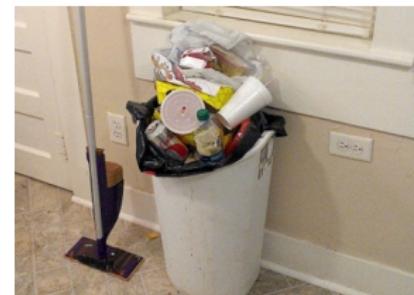
Reciprocal Free Riding



Reciprocal Free Riding



Reciprocal Free Riding



Recall the Ross & Sicoly (1979) housework study?

Negative Reciprocity

- What if someone is watching?
- The impact of negative reciprocity changes radically if subjects are given the opportunity to observe the contributions of others
 - and to punish those who do not contribute

Alternative Public Goods Game

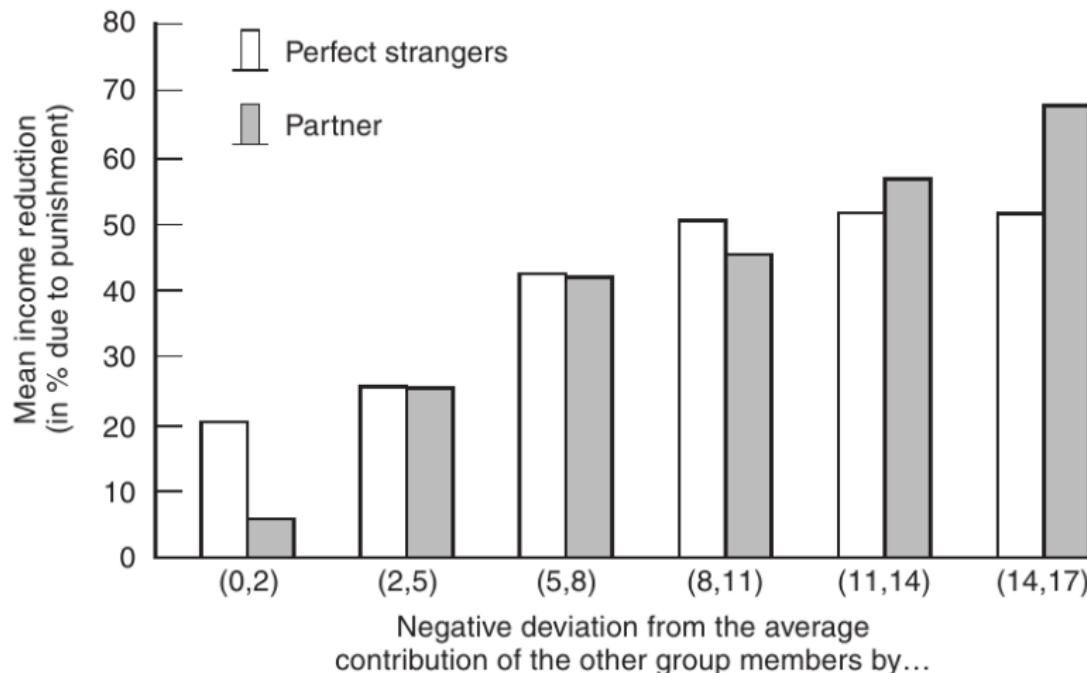
- Suppose each subject in a group has the opportunity to reduce the income of each other subject in the group
- e.g., reduction of the income of one other group member by x tokens
- Punishing costs $\frac{1}{3} \times x$
 - Which means selfish subjects will never punish (If all subjects were purely self-interested, contributions would be unaffected by the punishment opportunity)
 - Negatively reciprocal type subjects, will use the costly punishment opportunity to punish free riders
- The public good game with direct punishment provides an opportunity for reciprocal types to induce the selfish types to make “cooperative” choices

Public Goods and Punishment

- Same structure as previous 4 person 20 token game
- 2 versions:
 - “Perfect stranger”
 - Participant groups *shuffled* after each trial
 - 6 groups of 4 completed 6 trials
 - Nobody met another group member more than once
 - “Partner”
 - Same 4 group members across 10 trials
- Opportunity to **punish** after each trial (no punishment in control groups) (Fehr & Gächter, 2000)

Public Goods and Punishment: Results

Public Goods and Punishment: Results

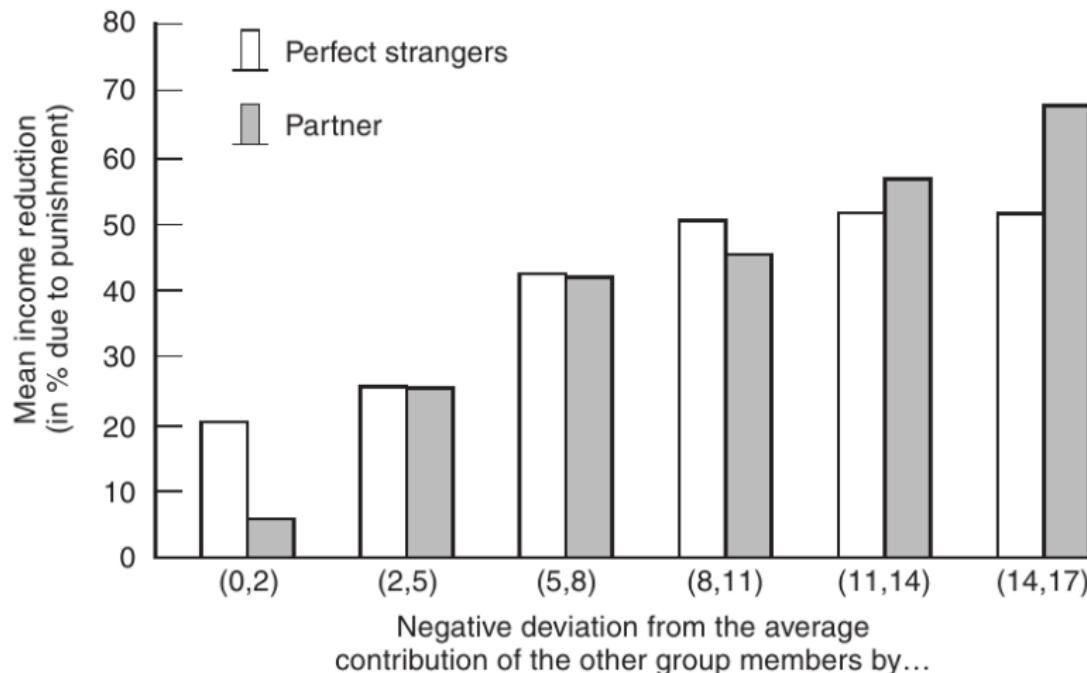


Mean income reduction for a given negative deviation from the mean contribution of other group members. Source: Fehr and

Public Goods and Punishment: Results

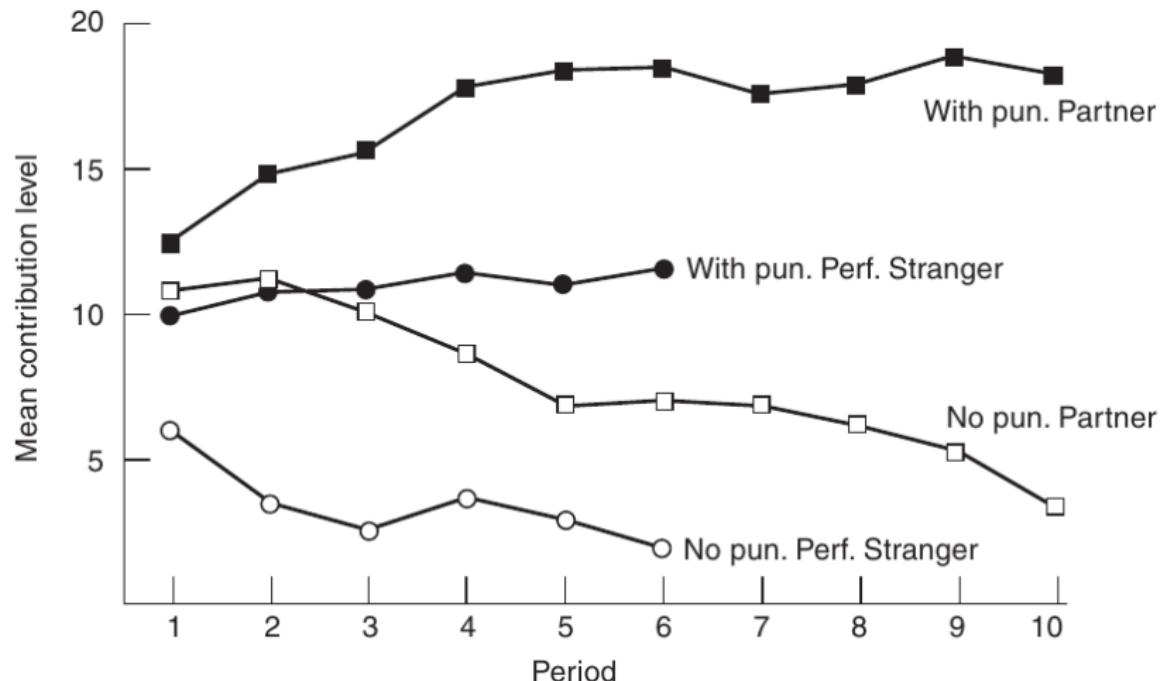
- Negative deviations from the Mean contributions of other group members was punished
- The larger the deviation the larger the punishment
- The more a subject free rides relative to the others the more it gets punished
- Moreover, this pattern is almost the same in the two versions of the game:
 - Free riders are punished irrespective of whether there are future rewards for the punisher
- Questionnaire evidence suggests that the deviation from the norm of cooperation causes resentment and the impulse to punish

Public Goods and Punishment: Results



Mean income reduction for a given negative deviation from the mean contribution of other group members. Source: Fehr and

Public Goods and Punishment: Results

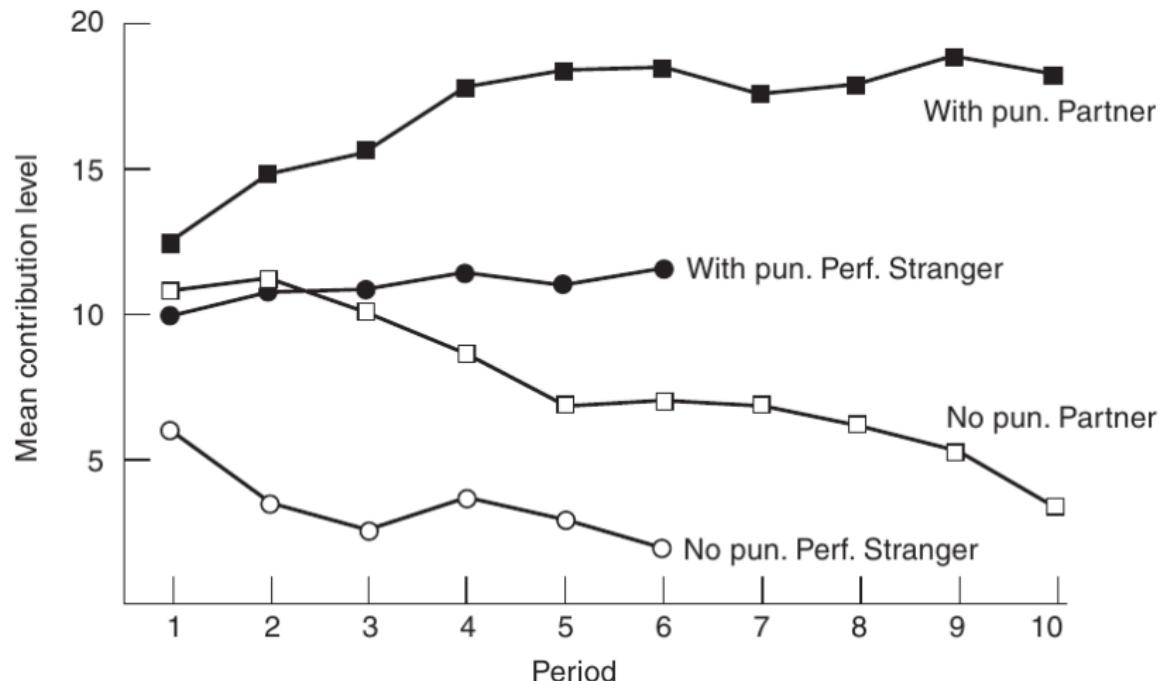


Evolution of average contributions with and without the punishment option in the Partner and the Perfect Stranger

Public Goods and Punishment: Results

- Absence of punishment leads to smaller contributions
 - (As predicted)
- Punishment leads to higher contributions

Public Goods and Punishment: Results



Evolution of average contributions with and without the punishment option in the Partner and the Perfect Stranger

Punishment and 'Free-riding'

<https://www.youtube.com/embed/8SOQduoLgRw?start=622>

Jonathan Haidt discussing Fehr & Gächter (2002)

Social Norms

From Public Goods to Social Norms

- Failure to behave “fairly” results in punishment
 - Where behaving fairly is the **norm**
- Behaviours that deviate from the **norm** are punished (sanctioned)
 - Sanctioning is often costly

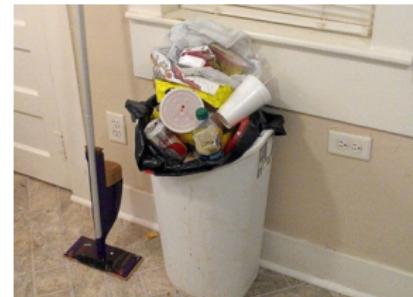
Sanctioning and Norms



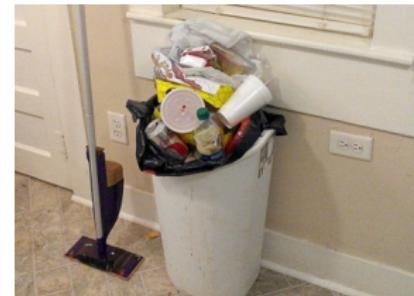
Sanctioning and Norms



Sanctioning can be Costly



Sanctioning can be Costly



Which is more costly? The sanction or the consequence of not sanctioning

Norms

- Norms are pervasive in social and economic life
- The large majority of interactions in people's lives are not regulated by explicit contracts
 - but by informal social norms.
- Norm-governed attitudes, social interactions, and conformism among peers, among relatives, and in neighborhoods
 - May have important consequences for human capital decisions
 - The decision to take part in elections
 - Criminal activities
 - Tax evasion
 - Abuse of welfare payments

Third Party Punishment

Third Party Punishment

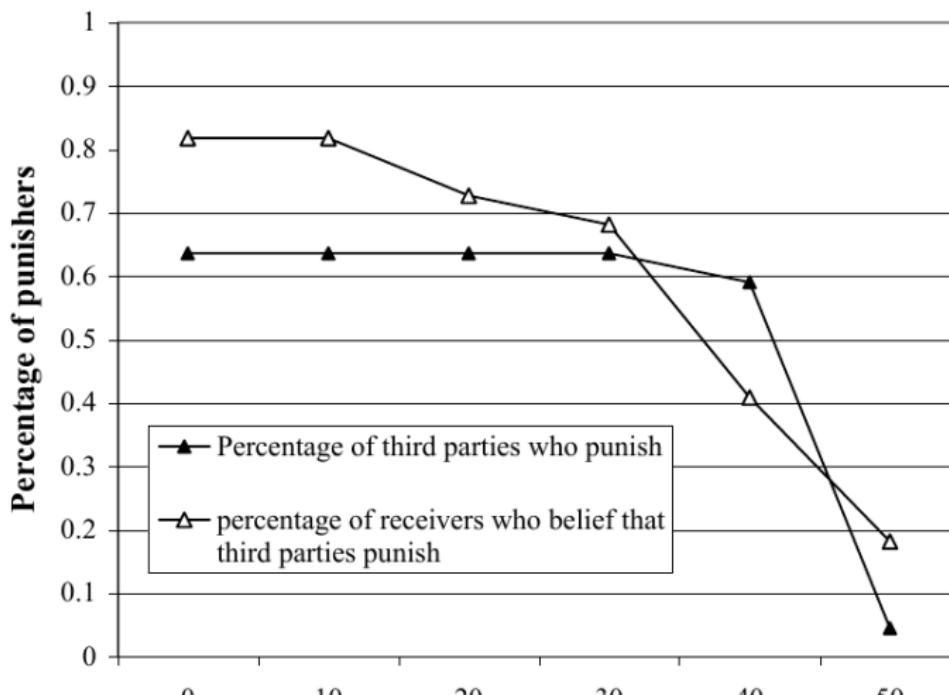
- Third Party Punishment in the Dictator game
- Players **A** (Dictator), **B** (Recipient) and **C** (Third Party)
- **A** has 100 points
 - Can give 0, 10, 20, 30, 40, or 50 points to **B**
- **B** has nothing
- **C** (has 50 points) can sanction **A**
 - assigning a sanction “costs” points
 - e.g., **C** spends 1 point to deduct 3 points from **A**(Fehr & Fischbacher, 2004)

Third Party Punishment

- Variables of Interest:
 - Frequency of punishment
 - Both *actual* and *expected*
 - Severity of punishment
 - Both *actual* and *expected*
 - Level/value of transfer (from dictators)
 - Both *actual* and *expected*

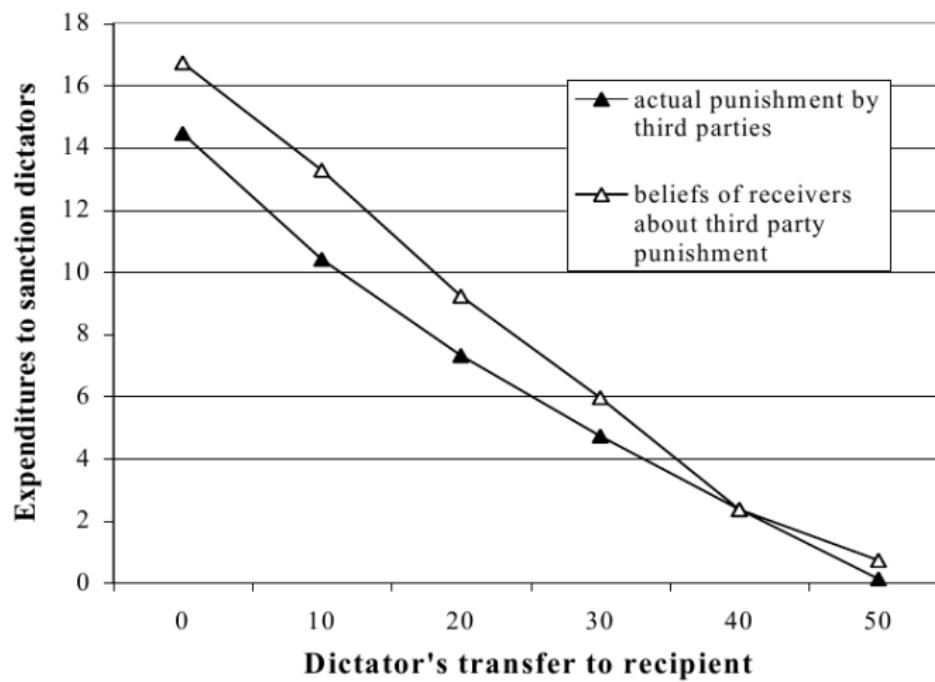
Third Party Punishment

Figure 1: Percentage of third parties who punish in the dictator game



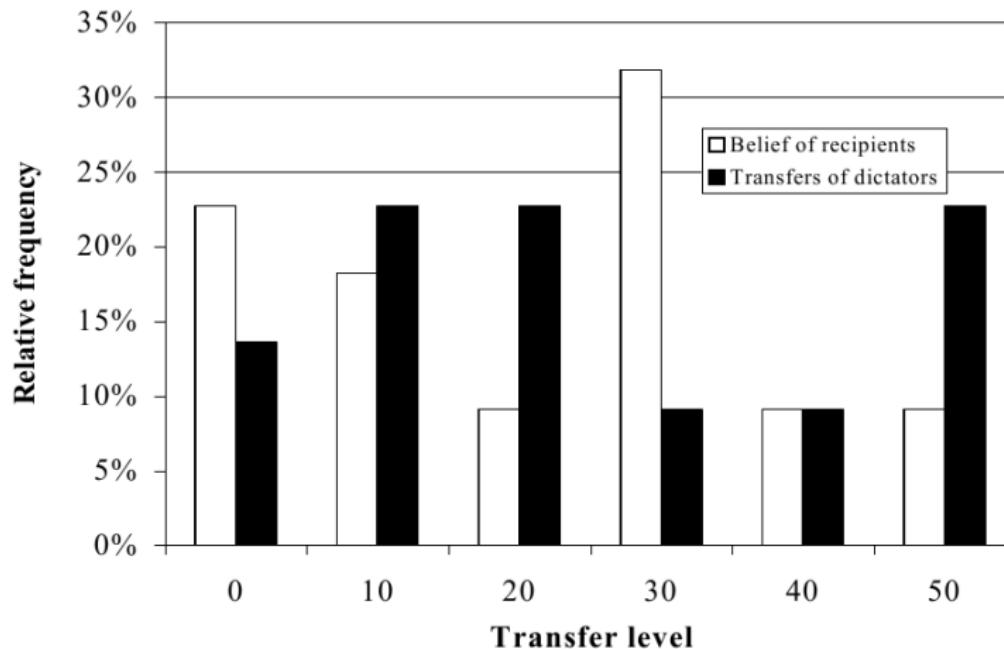
Third Party Punishment

Figure 2: Pattern of third party punishment in the dictator game



Third Party Punishment

Figure 3: Distribution of actual and expected transfers in the dictator game

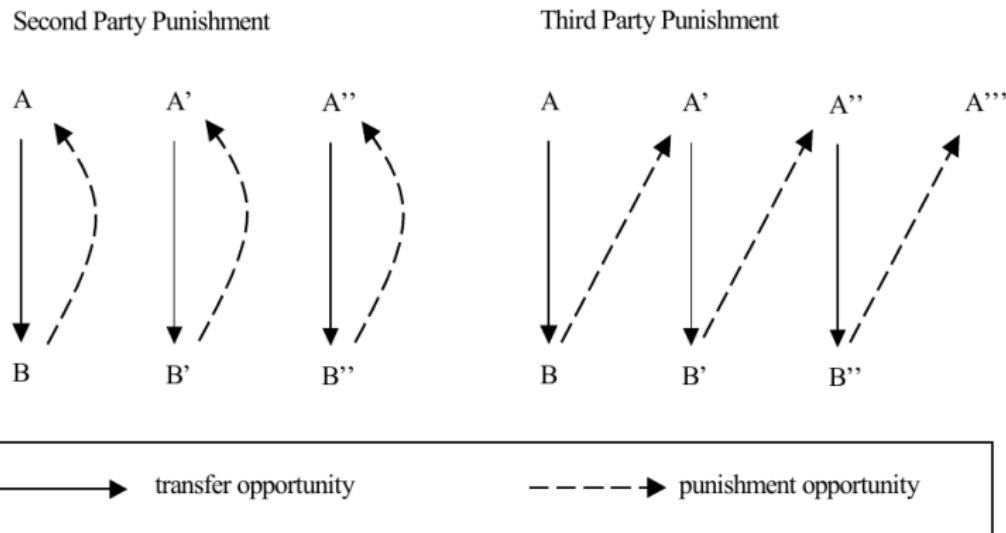


Punishment Expectation

- Recipients' beliefs about punishment appear to be higher than the actual punishment
- Recipients overestimate both the strength and frequency of punishment
- For each reduction of the transfer by 10 units, the recipients expect that the third parties will reduce the dictators' income by 9.9 units

Second and Third Party Punishment

Figure 4: The comparison of second and third party punishment in the dictator game.



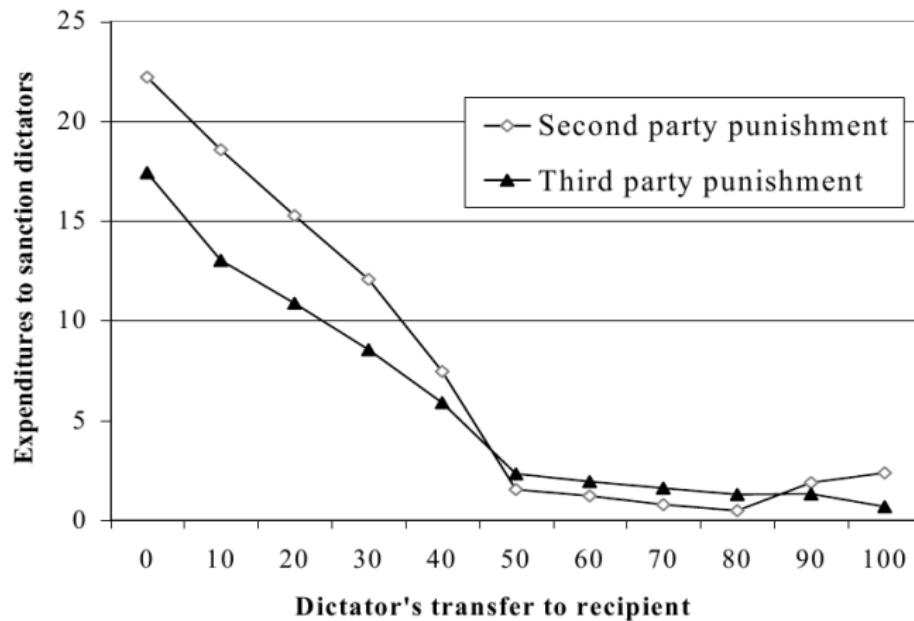
(Fehr & Fischbacher, 2004, p. 75)

Second and Third Party Punishment

- Variables of Interest:
 - Expenditure to sanction dictator
 - Both *second* and *third* party
 - Expected payoff for dictator
 - Both *second* and *third* party

Second and Third Party Punishment

Figure 5: Comparison of second and third party punishment

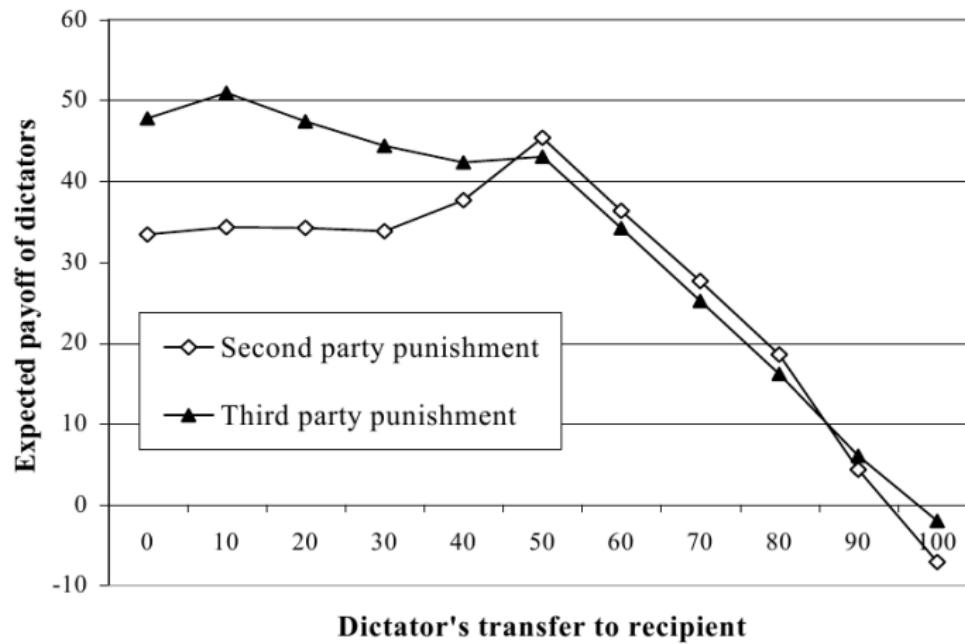


Second and Third Party Punishment

- Dictators face severe sanctions in the second and third party condition
- Second party sanctions for transfers below the egalitarian level are considerably stronger than those of the third party
- Low transfers are profitable for the dictators in the TP condition but not in the SP condition

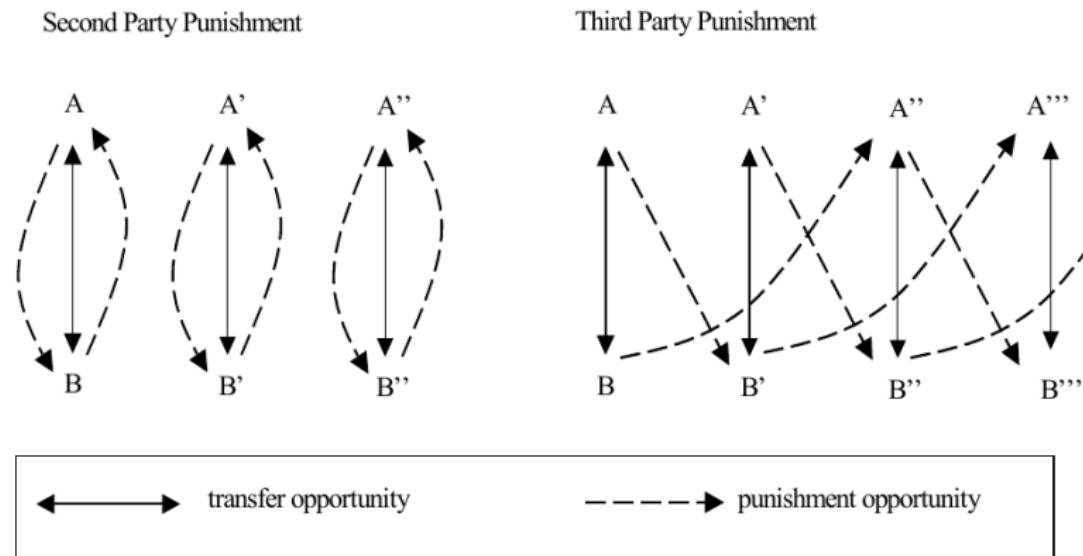
Second and Third Party Punishment

Figure 6: Expected payoffs of dictators under second and third party punishment



Third Party Punishment

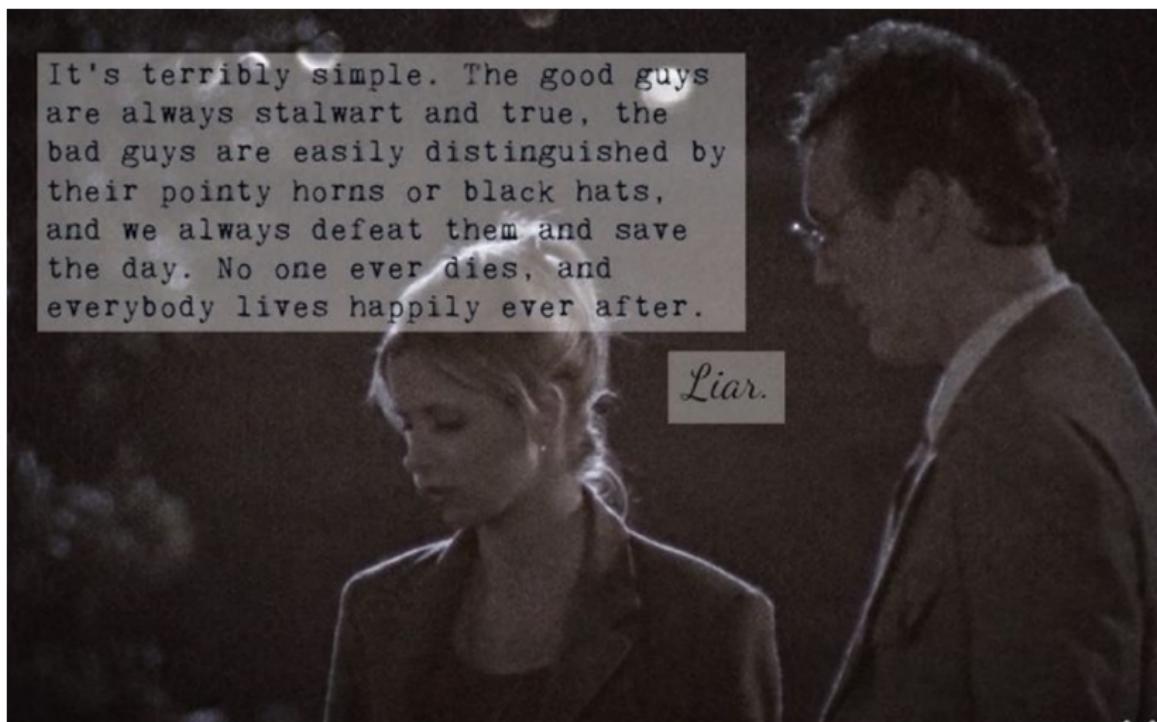
Figure 7: The comparison of second and third party punishment in the prisoners' dilemma.



Moral Judgments

<https://www.youtube.com/embed/wEjGVGtd2-Y?si=268c-IAGp9GuxEiM&start=90>

Moral Judgments



Known Influences on Moral Judgements

- Emotional influences (e.g., C. D. Cameron, Payne, & Doris, 2013)
- Intentionality, Evitability, benefit recipient (Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Christensen & Gomila, 2012)
- Action-outcome distinction (Crockett, 2013; Cushman, 2013)
- Trustworthiness and social evaluation (Everett, Faber, Savulescu, & Crockett, 2018; Everett, Pizarro, & Crockett, 2016)
- Personal-impersonal distinction (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001)
- Doctrine of double effect (Mikhail, 2000)
- Level of physical contact (Valdesolo & DeSteno, 2006)
- Order effects (Wiegmann, Okan, & Nagel, 2012)

Understanding this complexity

- **Moral Emotions:** Prinz (2005)
 - ‘*Emotions, I will suggest are perceptions of our bodily states. To recognise the moral value of an event is, thus, to perceive the perturbation that it causes*’ (Prinz, 2005, p. 99)
- **Social Intuitionist Model** (Haidt, 2001)
 - e.g., intuitions as phonemes (Haidt, 2001, p. 827)
 - ‘brain has a kind of gauge’, a ‘like-ometer’ (Haidt & Björklund, 2008, p. 187)
- **Theory of Dyadic Morality:** (Gray, Young, & Waytz, 2012; Schein & Gray, 2018)
 - Content focused (rather than underlying processes)
- **Dual Process Theories:** (e.g., Greene, 2008)
 - ‘Emotion’ vs ‘Cognition’ (not well defined)
- **Categorization** (McHugh, McGann, Igou, & Kinsella, 2022)

Theory of Dyadic Morality (TDM)

- “an intentional agent causing damage to a vulnerable patient” (Schein & Gray, 2018, p. 33)
- “an intentional moral agent and a suffering moral patient (Gray et al., 2012, p. 101).
- “interpersonal harm is the essence of morality” (Gray et al., 2012, p. 109)



Issues with TDM

Harmless Wrongs: Moral Dumbfounding

- Moral dumbfounding occurs when people defend a moral judgement even though they cannot provide supporting reasons

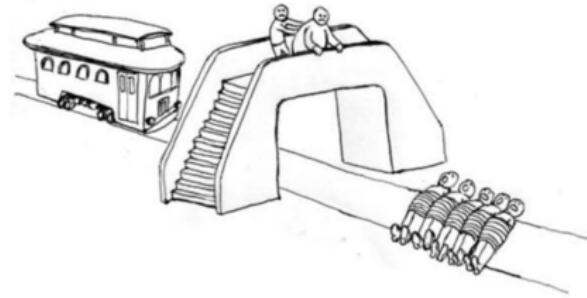
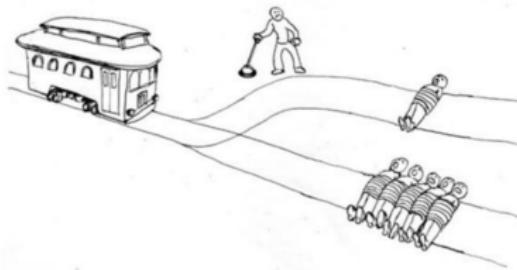
(Haidt, Björklund, & Murphy, 2000; McHugh, McGann, Igou, & Kinsella, 2017; McHugh, Zhang, Karnataka, Lamba, & Khokhlova, 2023)

The screenshot shows a research article from the journal *Memory & Cognition*. The article is titled "Searching for Moral Dumbfounding: Identifying Measurable Indicators of Moral Dumbfounding" and is authored by Cillian McHugh¹, Marek McGann¹, Eric R. Igou¹ and Elaine L. Kinsella¹. The article was published on December 11, 2022, and is available at <https://doi.org/10.3758/MC.40.1.421>. The abstract discusses moral dumbfounding, noting that it occurs when people defend a moral judgment even though they cannot provide a reason for this judgment. The phenomenon has been studied in WEIRD populations, and the current study extends this to non-Western samples. The article includes sections on methodology, results, and conclusions, along with a figure showing correlations between various measures.

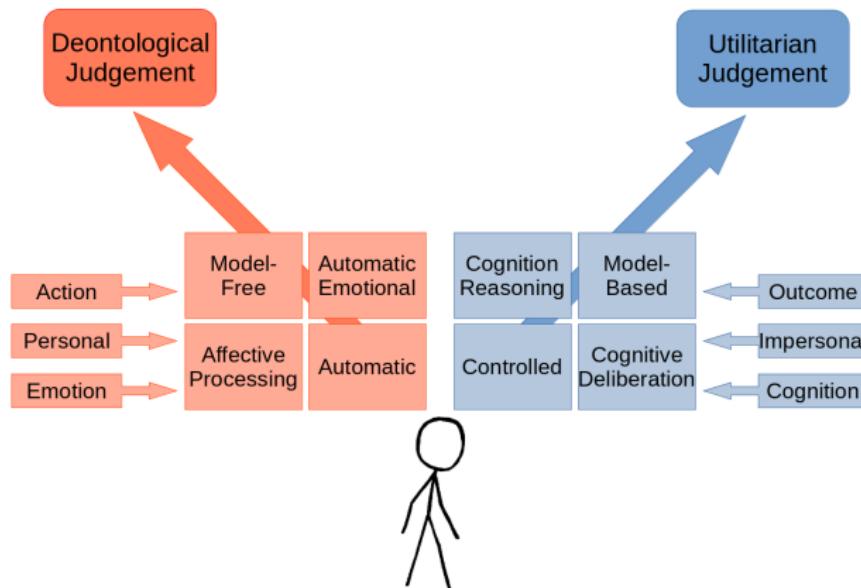
Moral Dumbfounding

Jennifer works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research. One night Jennifer is leaving the lab when she sees a body that is going to be discarded the next day. Jennifer was a vegetarian, for moral reasons. She thought it was wrong to kill animals for food. But then, when she saw a body about to be cremated, she thought it was irrational to waste perfectly edible meat. So she cut off a piece of flesh, and took it home and cooked it. The person had died recently of a heart attack, and she cooked the meat thoroughly, so there was no risk of disease.

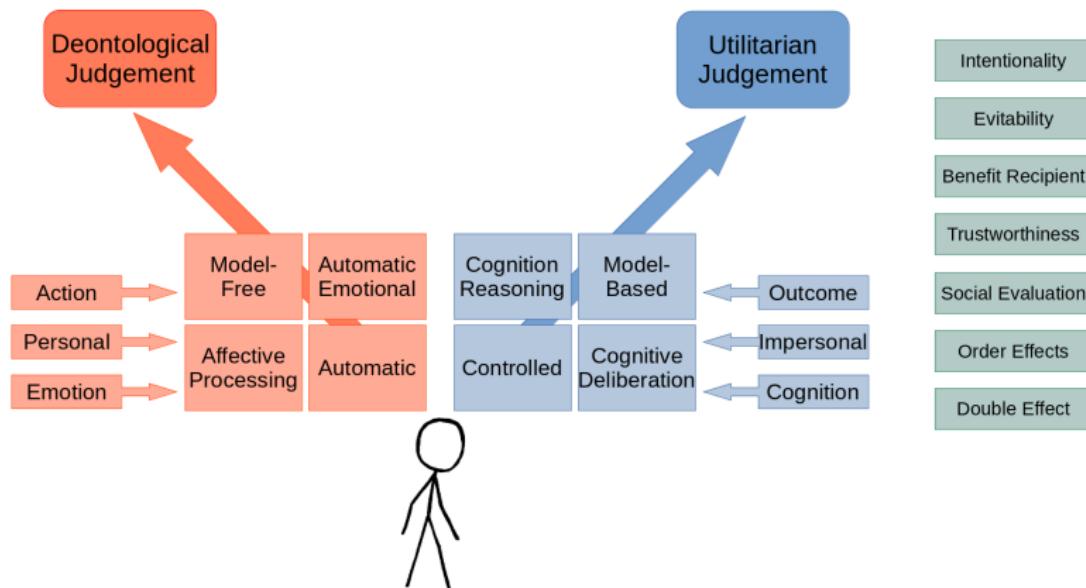
Dual-Process Approaches



Dual-Process Approaches



Dual-Process Approaches



Moral Judgment as Categorization

(MJAC)

Premises

- Moral judgment is a process of **categorizing** as *MORALLY RIGHT / MORALLY WRONG*
- **Domain general**
- **Context sensitive & Dynamical,**
Goal-Directed Activity

Core Predictions

- *Stability* emerges through continued and consistent type-token interpretation
- *Robustness* emerges through consistency across multiple contexts

Moral Judgment as Categorization (MJAC)

Cillian Mottaghi^{1,2,3,4}, Mark McCleary⁵, Eric R. Igou^{1,2,3},
and Elmer L. Kinchla^{2,3,4}
¹Department of Psychology, University of Colorado Boulder, Boulder, CO, USA;
²Department of Psychology, Penn State University, University Park, PA, USA;
³Department of Psychology, University of Colorado Denver, Denver, CO, USA;
⁴University of Colorado and Western Air Defense, Fort Carson, CO, USA



Psychological Science
Volume 30 Number 10 October 2019
ISSN: 0898-2603
eISSN: 1943-3584
DOI: 10.1137/18-0303
Copyright © 2019 by the American Psychological Society
0898-2603/19/3010-0001-12\$15.00
https://doi.org/10.1137/18-0303

Abstract
Moral variability and complexity of judgments of "right" and "wrong" cannot be easily accounted for within traditional approaches to moral judgment. Moral judgment as categorization (MJAC) leverages principles of categorization from cognitive science to account for the stability and robustness of moral judgments. MJAC predicts that people are more likely to make context-sensitive categorizations. They know that various actions (crosses between people, etc.) can be right or wrong depending on the context in which they occur. MJAC also predicts that people are more likely to make moral judgments according to the self-direction account of moral categorization. The learning and the habituation of the learning of moral categories are also predicted by MJAC. MJAC also predicts that people are more likely to make moral judgments for the complexity of moral judgments. MJAC offers greater explanatory power than existing approaches while also accounting for the stability and robustness of moral judgments.

Keywords

moral, categorization, cognitive function, moral judgment
In words simple, the predictions are intuitive and easy. The findings are easily distinguished by other theories. MJAC is a theory that is both simple and useful. It can be used to predict and to describe them and use the theory. Nobody ever does anything more useful than this.

or apparent inconsistencies in judgments have been addressed in terms of people's ability to do a specific task. MJAC is a theory that is both simple and useful. It can be used to predict and to describe them and use the theory. Nobody ever does anything more useful than this.

Ad-Hoc Goal-Derived Categories



- *THINGS TO PACK INTO A SUITCASE* (Barsalou, 1991, 2003)
- Type-token interpretation

The Organisation of Categories: Fruit

The Organisation of Categories: Fruit



The Organisation of Categories: Fruit



The Organisation of Categories: Fruit



The Organisation of Categories: Fruit



Evidence

Phenomenon		Categorisation	Morality
Variability	Interpersonal	✓	✓
	Intrapersonal	✓	✓
Context	Culture	✓	✓
	Social	✓	✓
	Development	✓	✓
	Emotion	✓	✓
	Framing	✓	✓
	Language	✓	✓
	Order/recency	✓	✓
	Skill	✓	✓
Other	Typicality	✓	Hypothesised
	Dumbfounding	✓*	✓

* Not explicitly discussed as “dumbfounding” however there are key similarities that suggest parity

Similarity of Explanation

Morality

Order Effects:

Petrinovich & O'Neill
(1996)

Categorization

Barsalou (1982), Higgins, Bargh, &
Lombardi (1985)

Language Effects:

Cipolletti, McFarlane, &
Weissglass (2016)

Colbeck & Bowers (2012), Harris,
Ayçiçeği, & Gleason (2003)

Emotion Effects:

C. D. Cameron et al.
(2013)

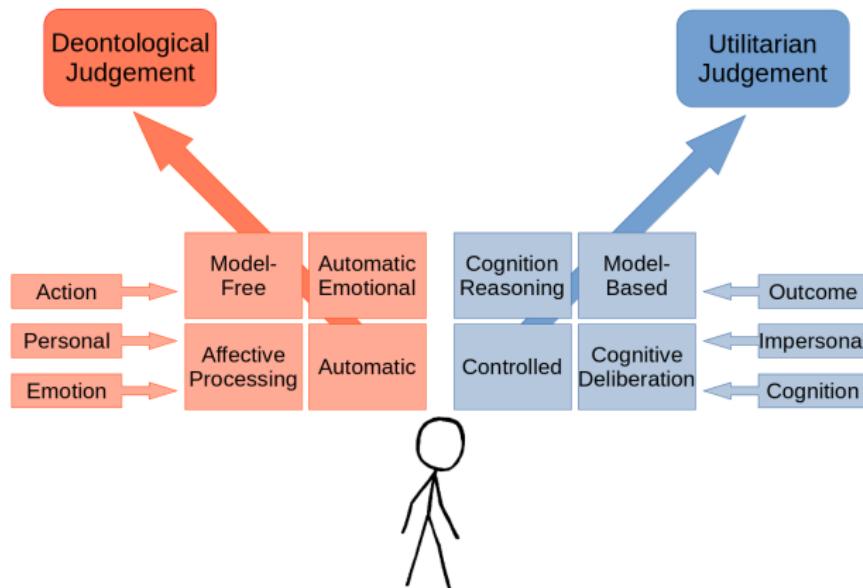
Barsalou & Wiemer-Hastings (2005)

Typicality:

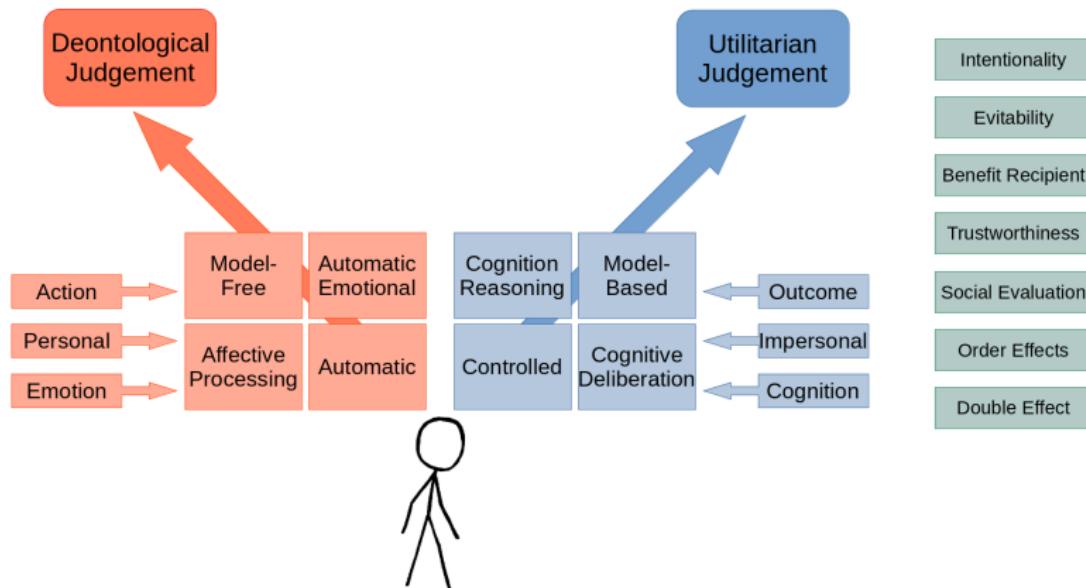
Gray & Keeney (2015)

McCloskey & Glucksberg (1978), Oden
(1977)

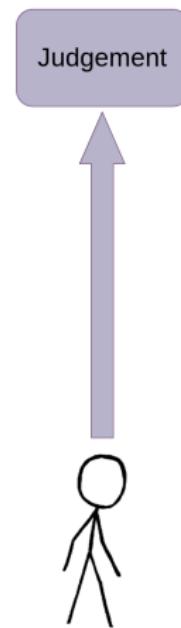
Contrasting with Existing Approaches



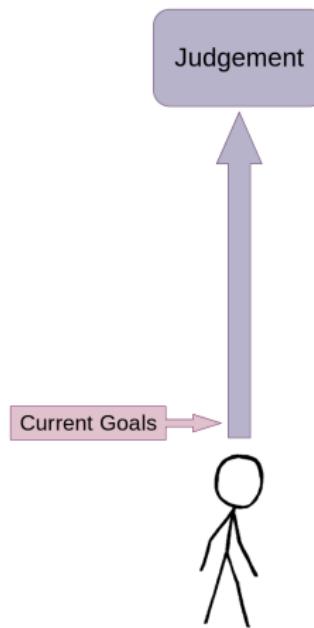
Contrasting with Existing Approaches



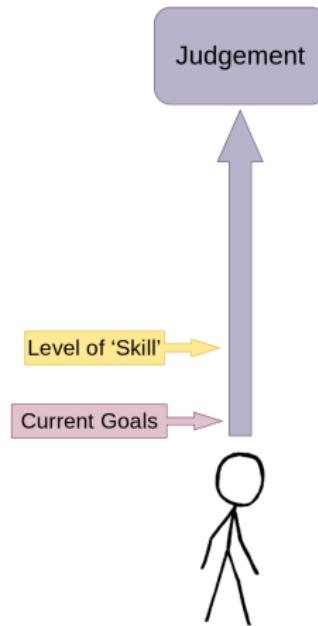
MJAC



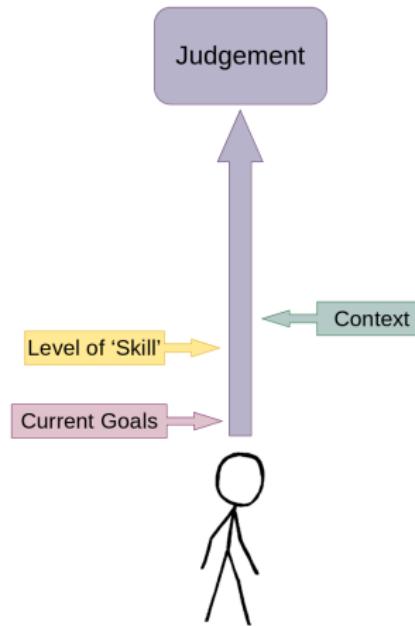
MJAC



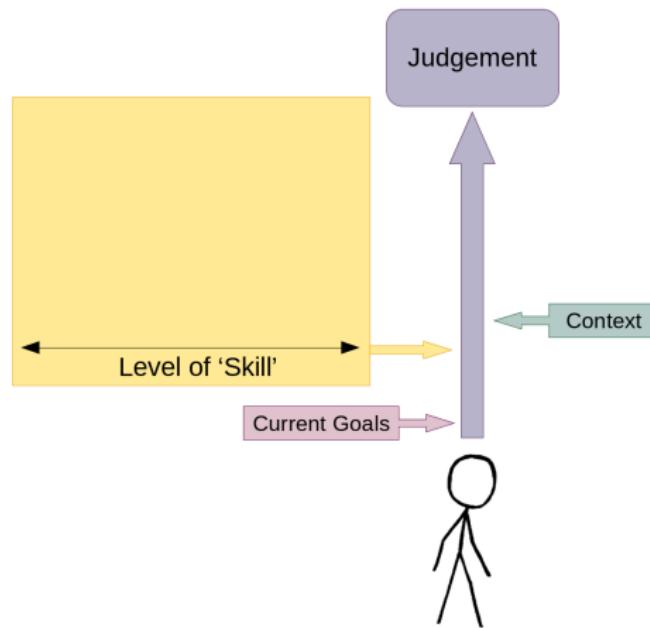
MJAC



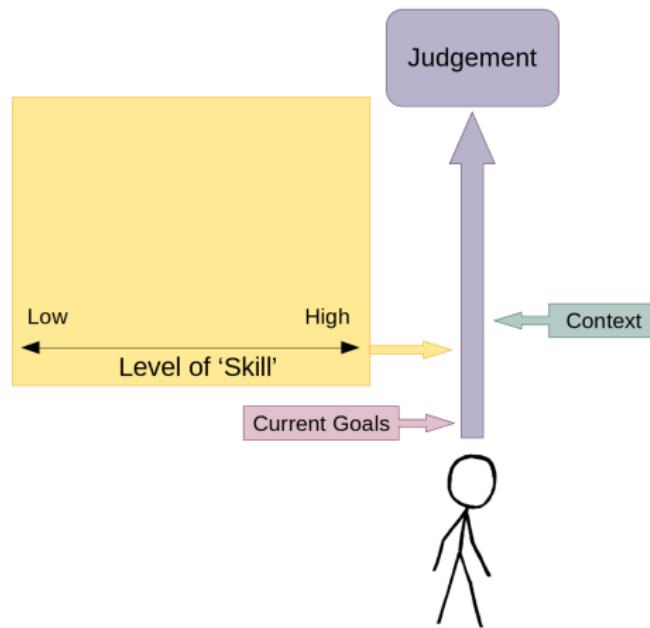
MJAC



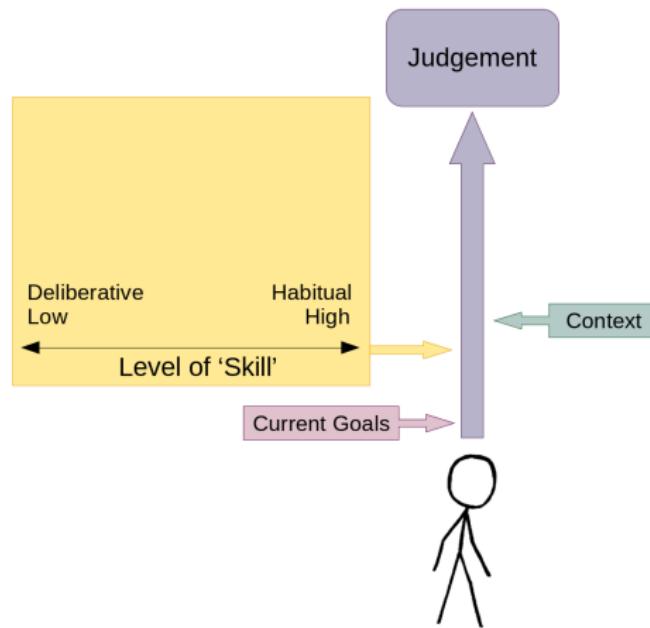
MJAC



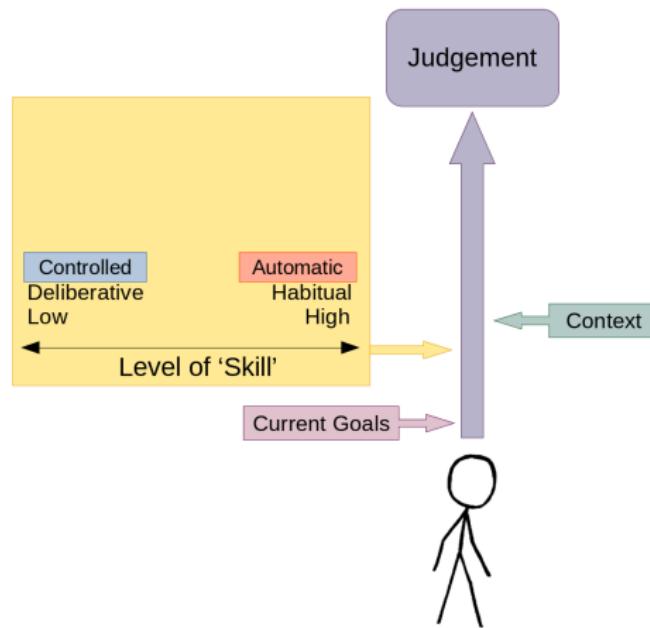
MJAC



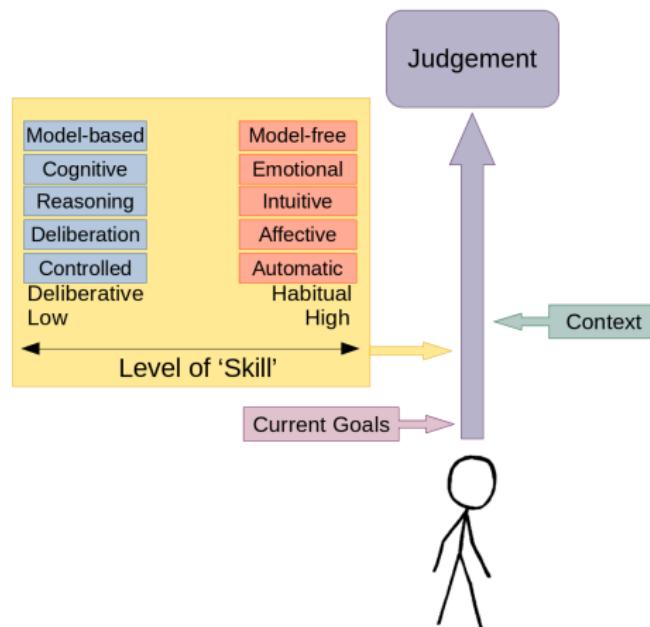
MJAC



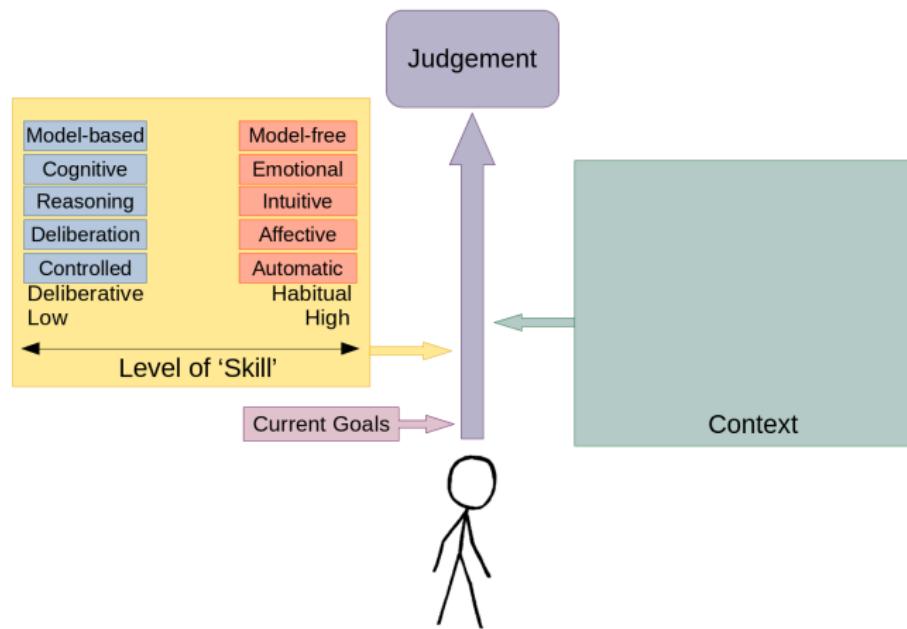
MJAC



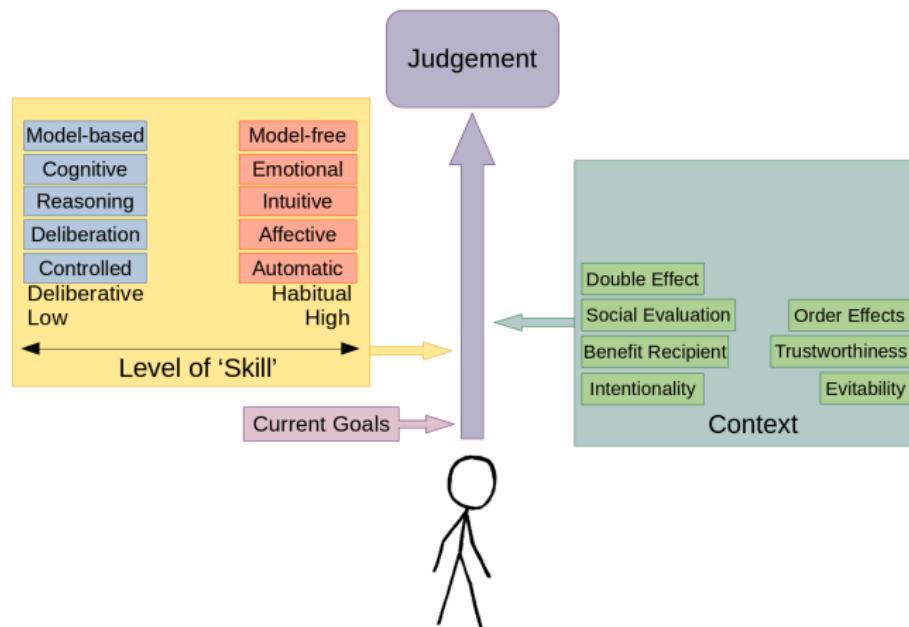
MJAC



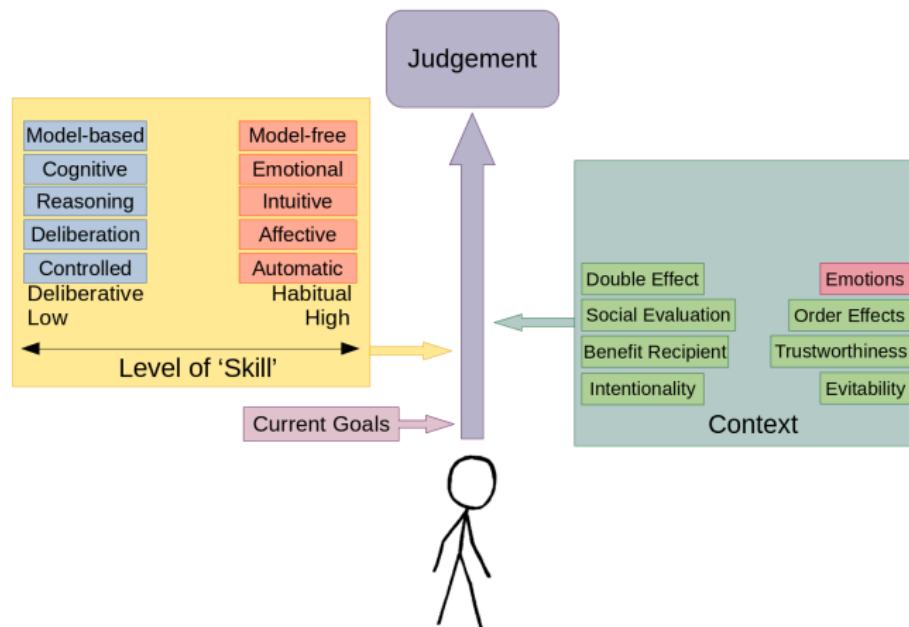
MJAC



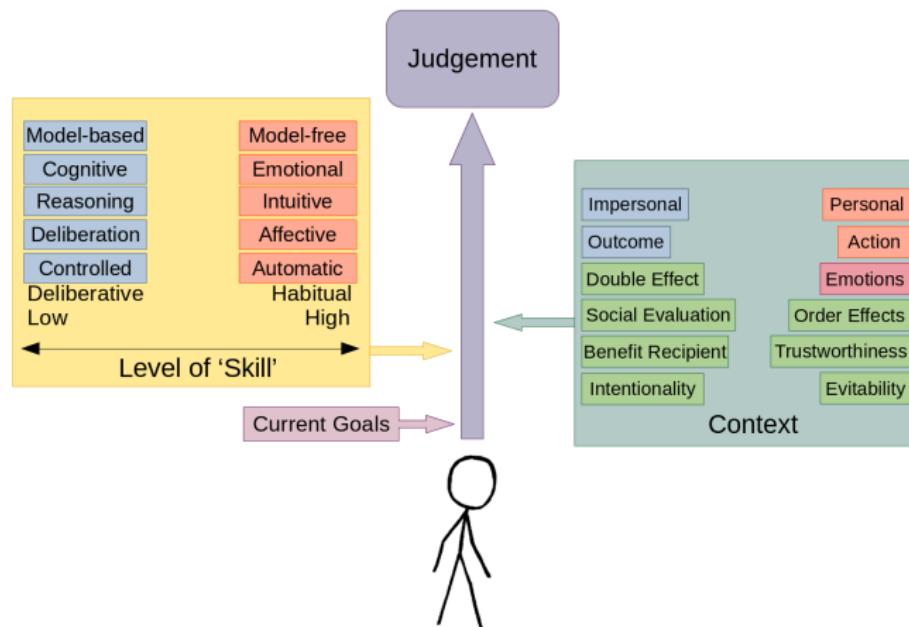
MJAC



MJAC



MJAC



In-Class Activity

In-class Activity

- Discuss the following:
- Can you identify any other methods for studying the influence of fairness and ethics in decision making?

References

References

- Abbink, K., Irlenbusch, B., & Renner, E. (2000). The moonlighting game: An experimental study on reciprocity and retribution. *Journal of Economic Behavior & Organization*, 42(2), 265–277. [https://doi.org/10.1016/S0167-2681\(00\)00089-5](https://doi.org/10.1016/S0167-2681(00)00089-5)
- Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, 10(1), 82–93. <https://doi.org/10.3758/BF03197629>
- Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 76–121). San Diego: Academic Press.
- Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18(5-6), 513–562. <https://doi.org/10.1080/01690960344000026>